A BAYESIAN APPROACH TO SOME MISSING VALUE PROBLEMS IN ANOVA AND CONTINGENCY TABLES

by

Daniel Bloch

Technical Report No. 62

November, 1966

*This research was sponsored by the Office of Naval Research on Contract Nonr 4010(09) awarded to the Department of Statistics The Johns Hopkins University.

DEPARTMENT OF STATISTICS

THE JOHNS HOPKINS UNIVERSITY

BALTIMORE, MARYLAND

A Bayesian Approach to some Missing Value Problems in ANOVA and Contingency Tables*

by

Daniel Bloch

1. Introduction

Non-Bayesian procedures for dealing with missing values in ANOVA and contingency tables are well-known (see e.g. [3] and [6]). In this paper we derive the appropriate Bayesian procedures for a randomized block design and for an r × c contingency table. The posterior distributions for the missing observation are also given. This paper shows that the classical non-Bayesian procedures do have a simple and natural Bayesian interpretation.

2. Missing values in a randomized block design.

Suppose that the observations in a b-block × t-treatment randomized block design are incomplete because y_{11} is missing. Our model for the available observations is

$$y_{ij} = \mu + \tau_j + \beta_i + \epsilon_{ij} \ , \ (i, j) \neq (1, 1); \ i=1,\dots,b \ ; \ j=1,\dots,t. \tag{2.1}$$

We assume that $\sum_{i=1}^{b} \beta_i = 0$, $\sum_{j=1}^{t} \tau_j = 0$, and the ϵ_{ij}'s are independent and normally distributed with mean zero and variance σ^2. (μ is the over-all mean effect, τ_j $(j=1,\dots,t)$ is the j^{th} treatment effect, and $\beta_i(i=1,\dots,b)$ is the i^{th} block effect).

*This research was sponsored by the Office of Naval Research on Contract Nonr 4010(09) awarded to the Department of Statistics The Johns Hopkins University.

We now make the usual "non-informative" assumption that the joint prior distribution of (μ, σ, τ_j's, β_i's) is proportional to $1/\sigma$. From the resulting posterior distribution, μ, σ and the β_i's may be integrated out to obtain the distribution of most interest--the posterior distribution of the τ_j's. The lack of balance caused by the missing observation y_{11} makes this an awkward calculation. If however we introduce y_{11} as a random variable, this balance is restored. Since $y_{11} = \mu + \tau_1 + \beta_1 + \epsilon_1$, the density of y_{11} given $\mu + \tau_1 + \beta_1$ is normal wit.. mean $\mu + \tau_1 + \beta_1$ and variance σ^2. The joint posterior density of (y_{11}, μ, σ, τ_j's, β_i's) is therefore proportional to the product of the conditional normal density of y_{11} and the lik...ihood of the available observations and σ^{-1}. Hence

$$\text{Post }(\mu,\sigma,\tau_j\text{'s},\beta_i\text{'s},y_{11}) \propto \frac{1}{\sigma^2}\exp\left\{\frac{-1}{2\sigma^2}(y_{11}-\mu-\tau_1-\beta_1)^2\right\}\times \begin{array}{l}\text{Likelihood function}\\ \text{for the available}\\ \text{observations}\end{array}$$

$$\propto \frac{1}{\sigma^{tb+1}}\exp\left\{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{b}\sum_{j=1}^{t}(y_{ij}-\mu-\tau_j-\beta_i)^2\right]\right\} . \tag{2.2}$$

The exponent can be written as

$$\sum_{i=1}^{b}\sum_{j=1}^{t}(y_{ij}-\mu-\tau_j-\beta_i)^2 = tb(\bar{y}-\mu)+b\sum_{j=1}^{t}(\bar{y}_j-\bar{y}-\tau_j)^2 + t\sum_{i=1}^{b}(\bar{y}_i-\bar{y}-\beta_i)^2 + S^2, \tag{2.3}$$

where

$$y_{i.} = \sum_{j=1}^{t} \frac{y_{ij}}{t}, \quad \bar{y}_{.j} = \sum_{i=1}^{b} \frac{y_{ij}}{b}, \quad \bar{y} = \sum_{i=1}^{b} \sum_{j=1}^{t} \frac{y_{ij}}{tb}, \text{ and}$$

$$S^2 = \sum_{i=1}^{b} \sum_{j=1}^{t} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$$

Substituting (2.3) into (2.2) we have that

$$\text{Post }(\sigma, \tau_j\text{'s}, y_{11}) \propto \frac{1}{\sigma^{tb-b+1}} \exp\left\{\frac{-1}{2\sigma^2}\left[S^2 + b\sum_{j=1}^{t} (\bar{y}_{.j} - \bar{y} - \tau_j)^2\right]\right\} \qquad (2.4)$$

$$\times \int \frac{d\mu}{\sigma} \exp\left\{\frac{-tb}{2\sigma^2}(\bar{y} - \mu)^2\right\} \quad \int \frac{d\beta_1 \ldots d\beta_b}{\sigma^{b-1}} \exp\left\{\frac{-t}{2\sigma^2}\sum_{i=1}^{b} (\bar{y}_i - \bar{y} - \beta_i)^2\right\}$$

(b-1)-dimensional
space with $\sum_{i=1}^{b} \beta_i = 0$

$$\propto \frac{1}{\sigma^{tb-b+1}} \exp\left\{\frac{-1}{2\sigma^2}\left[S^2 + b\sum_{j=1}^{t} (\bar{y}_{.j} - \bar{y} - \tau_j)^2\right]\right\}.$$

Denote the missing observation, y_{11}, by x.

Let $\bar{y}' = \bar{y} - \frac{x}{tb}$, $\bar{y}'_{1.} = \bar{y}_{1.} - \frac{x}{t}$, $\bar{y}'_{.1} = \bar{y}_{.1} - \frac{x}{b}$.

It is easily verified that

$$b\sum_{j=1}^{t}(\bar{y}_{.j}-\bar{y}-\tau_j)^2 = \frac{x^2(t-1)}{tb} - 2x(\bar{y}'-\bar{y}'_{.1}+\tau_1)$$

$$+\, b\sum_{j=2}^{t}(\bar{y}_{.j}-\bar{y}'-\tau_j)^2 + b(y'_{.1}-y'-\tau_1)^2 \qquad , \tag{2.5}$$

and

$$S^2 = \sum_{i=1}^{b}\sum_{j=1}^{t}{}^{/}\,(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}')^2 + \frac{x^2(t-1)(b-1)}{tb}$$

$$+\, 2x(\bar{y}'-\bar{y}'_{1.}-\bar{y}'_{.1}) + (\bar{y}'-\bar{y}'_{1.}-\bar{y}'_{.1})^2 \quad , \tag{2.6}$$

where "/" means $\bar{y}_{1.}$ and $\bar{y}_{.1}$ are to be replaced by $\bar{y}'_{1.}$ and $\bar{y}'_{.1}$ respectively and the summation does not incluse $(i,j) = (1,1)$, i.e.,

$$\sum_{i=1}^{b}\sum_{j=1}^{t}{}^{/}\,(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}-\bar{y}')^2 = \sum_{i=2}^{b}\sum_{j=2}^{t}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}')^2$$

$$+\sum_{j=2}^{t}(y_{ij}-\bar{y}'_{1.}-\bar{y}_{.j}+\bar{y}')^2 + \sum_{i=2}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}'_{.1}+\bar{y}')^2.$$

From (2.5) and (2.6) we have that the coefficients of the x^2 and $2x$ terms from $S^2 + b \sum_{j=1}^{t} (\bar{y}_{.j} - \bar{y} - \tau_j)^2$ are $\left(\frac{t-1}{t}\right)$ and $- (\bar{y}'_{1.} + \tau_1)^2$ respectively. Since

$$\left(\frac{t-1}{t}\right) x^2 - 2X(\bar{y}'_{1.} + \tau_1) = \left(\frac{t-1}{t}\right) \left[x - \frac{(\bar{y}'_{1.}+\tau_1)t}{t-1}\right]^2 - \frac{(\bar{y}'_{1.}+\tau_1)^2 t}{t-1}$$

we have that

$$\text{Post } (\sigma,\tau_j\text{'s}) \propto \frac{1}{\sigma^{tb-b}} \exp \left\{\frac{-1}{2\sigma^2}\left[\sum_{i,j} (\bar{y}_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}')^2 + (\bar{y}' - \bar{y}'_{1.}-\bar{y}'_{.1})^2 \right.\right. \tag{2.7}$$

$$\left.\left. + b\sum_{j=2}^{t} (\bar{y}_{.j}-\bar{y}'-\tau_j)^2 + b(\bar{y}'_{.1}-\bar{y}'-\tau_1)^2 - \frac{(\bar{y}'_{1.}+\tau_1)^2 t}{t-1}\right]\right\}.$$

If is easily shown that

$$(\bar{y}'-\bar{y}'_{1.}-\bar{y}'_{.1})^2 + b(\bar{y}'_{.1}-\bar{y}'-\tau_1)^2 - \frac{(\bar{y}'_{1.}+\tau_1)^2 t}{t-1} =$$

$$= \left(\frac{bt-b-t}{t-1}\right) \left[\tau_1 - \frac{\{b(t-1)(\bar{y}'_{.1}-\bar{y}')+t\bar{y}'_{1.}\}}{bt-b-t}\right]^2 - \frac{(b+t)}{bt-b-t} (\bar{y}'-\bar{y}'_{1.}-\bar{y}'_{.1})^2$$

Hence

$$\text{Post}\,(\tau_j\text{'s} \mid \text{available data}) \propto \frac{1}{[\,c^2+(\tau-a)'M(\tau-a)]^{\frac{tb-b-1}{2}}},\quad \sum_{j=1}^{t}\tau_j = 0, \tag{2.8}$$

where

$$a = \left(\frac{b(t-1)(\bar{y}'_{.1}-\bar{y}')+t\,\bar{y}'_{1.}}{bt-b-t},\ \bar{y}_{.2}-\bar{y}',\ldots,\bar{y}_{.t}-\bar{y}'\right)'$$

$$\tau = (\tau_1,\ldots,\tau_t),$$

$$\underset{t\times t}{M} = \begin{pmatrix} \frac{bt-b-t}{t-1} & & & \\ & b & & 0 \\ & & \ddots & \\ & 0 & & b \end{pmatrix},$$

and

$$c^2 = {\sum_{i,j}}'(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}')^2 - \frac{(b+t)}{bt-b-t}(\bar{y}'-\bar{y}_{1.}-\bar{y}'_{.1})^2$$

The posterior density, (2.8), is constant where $(\tau-a)'M(\tau-a)$ is constant. The surfaces $(\tau-a)'M(\tau-a) = c$ are ellipsoids in the $(t-1)$-dimensional spare with $\sum_{j=1}^{t}\tau_j = 0$. The density decreases as the distance from the center of the ellipsoids increases. Hence a confidence region is an ellipsoid. The same argument as used in proving theorem 6.4.1 in [5] shows that if

$$\phi = \frac{(\tau-a)'M(t-a)/(t-1)}{c^2/(t-1)(b-1)-1},$$

then

$$\text{Post }(\phi \mid \text{available observations}) \propto \frac{\phi^{\frac{t-3}{2}}}{[(t-1) + [(t-1)(b-1)-1]\phi^{\frac{tb-b-1}{2}}} \qquad (2.9)$$

and hence

$$\frac{(\tau-a)'M(\tau-a)/(t-1)}{C^2/(t-1)(b-1)-1} \sim F_{t-1,\ (t-1)(b-1)-1}. \qquad (2.10)$$

An α-level test of the null hypothesis that there are no treatment effects is therefore to declare the results significant if

$$\frac{a'Ma/(t-1)}{C^2/(t-1)(b-1)-1} > F^{(1-\alpha)}_{t-1,\ (t-1)(b-1)-1}$$

If no observations are missing, then it can be verified that the posterior distribution of $\phi_1 = \dfrac{(t-a_1)'M_1(\tau\cdot a_1)(t-1)}{S^2/(t-1)(b-1)}$ is $F_{t-1,(t-1)(b-1)}$

where $a_1 = (\bar{y}_{.1} - \bar{y}, \bar{y}_{.2} - \bar{y}, \ldots, \bar{y}_{.t}-\bar{y})$

and $\underset{t \times t}{M_1} = \begin{pmatrix} b & & O \\ & b & \\ & \ddots & \\ O & & b \end{pmatrix}$.

This is the same as the sampling result. The effect of the missing observation is to decrease the error degrees of freedom by unity. In the analysis a_1, M_1 and S^2 should be replaced by a, M and C^2 respectively.

We have from (2.4) that

$$\text{Post }(x,\sigma) \propto \frac{1}{\sigma^{tb-b-t+2}} \exp\left\{\frac{-S^2}{2\sigma^2}\right\} \int_{\substack{(t-1)\text{-dimensional}\\ \text{spare with } \sum\tau_j=0}} \frac{d\tau_1\ldots d\tau_t}{\sigma^{t-1}} \exp\left\{\frac{-1}{2\sigma^2}\left[b\sum_{j=1}^{t}(\bar{y}_{\cdot j}-\bar{y}-\tau_j)^2\right]\right\} \tag{2.11}$$

$$\propto \frac{1}{\sigma^{tb-b-t+2}} \exp\left\{\frac{-S^2}{2\sigma^2}\right\}$$

Hence, using the expression for S^2 given by (2.6),

$$\begin{aligned}\text{Post }(x,\sigma) \propto \frac{1}{\sigma^{tb-b-t-2}} \exp\Bigg\{\frac{-1}{2\sigma^2}\Bigg[&\sum_{i,j}{}'(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}')^2 \\ &+x^2\,\frac{(t-1)(b-1)}{tb}+2x(\bar{y}'-\bar{y}'_{1\cdot}-\bar{y}'_{\cdot 1}) \\ &+(\bar{y}'-\bar{y}'_{1\cdot}-\bar{y}'_{\cdot 1})^2\Bigg]\Bigg\}\end{aligned} \tag{2.12}$$

$$\begin{aligned}\propto \frac{1}{\sigma^{tb-b-t+2}} \exp\Bigg\{\frac{-1}{2\sigma^2}\Bigg[&\sum_{i,j}{}'(y_{ij}-\bar{y}_{\cdot j}-\bar{y}_{i\cdot}+\bar{y}')^2-\frac{(b+t-1)}{(t-1)(b-1)}(\bar{y}'-\bar{y}_{1\cdot}-\bar{y}'_{\cdot 1})^2 \\ &+\frac{(t-1)(b-1)}{tb}\left[x-\frac{(\bar{y}'_{1\cdot}+\bar{y}'_{\cdot 1}-\bar{y}')tb}{(t-1)(b-1)}\right]^2\Bigg]\Bigg\}\end{aligned}$$

upon completing the square with respect to x. Therefore the posterior distribution for the missing observation x is given by

$$\text{Post }(x) \propto \frac{1}{\left[D^2 + \frac{(t-1)(b-1)}{tb} \left(x - \frac{(\bar{y}'_{1 \cdot} + \bar{y}_{\cdot 1} - \bar{y}')tb}{(t-1)(b-1)} \right)^2 \right]} \quad \frac{tb+b-t+1}{2} \qquad (2.13)$$

where $D^2 = \sum_{i,j}{}' (y_{ij} - \bar{y}_{\cdot j} - \bar{y}_{i \cdot} + \bar{y}')^2 - \frac{(b+t-1)}{(t-1)(b-1)} (\bar{y}' - \bar{y}'_{1 \cdot} - \bar{y}'_{\cdot 1})^2$.

Notice that

$$\frac{(\bar{y}'_{1 \cdot} + \bar{y}'_{\cdot 1} - \bar{y}')tb}{(t-1)(b-1)} = \frac{bB + t\tau - b}{(t-1)(b-1)} = \text{yates estimator for the missing observation } x, \qquad (2.14)$$

where $G = t\, b\, \bar{y}' =$ Grand total of the available observations,

$B = t\, \bar{y}'_{1 \cdot} =$ total of the remaining units in the block where the missing unit appears,

$\tau = b\, \bar{y}'_{\cdot 1} =$ total of the yields of this treatment in the other blocks

From (2.13) we have that

$$\text{Post }(x) \propto \frac{1}{\left[1 + \frac{(t-1)(b-1)}{tb} \left(\frac{x - \mu(x)}{D} \right)^2 \right]^{\frac{(t-1)(b-1)}{2}}} \qquad (2.15)$$

where $\mu(x)$ is the Yates estimator for x and is given by (2.14).

If we let

$$\tau = \sqrt{\frac{(tb-b-t)(t-1)(b-1)}{tb}} \left(\frac{x-\mu(x)}{D} \right) ,$$

then

$$\text{Post }(v) \propto \frac{1}{\left[1 + \frac{v^2}{tb-b-t}\right]^{\frac{(t-1)(b-1)}{2}}} \quad ,\text{i.e.} \qquad (2.16)$$

v is distributed as a student-t random variable with $(tb-b-t) = (t-1)(b-1)-1$ degrees of freedom. It is interesting to note that if x is replaced by $\mu(x)$ in (2.6), then S^2 is identical to D^2. Inferences about the missing observation x should be made by referring to (2.16)

3. Missing values in contingency tables.

Let n_{ij} be the cell frequency in the $(ij)^{th}$ cell in an $r \times c$ contingency table. Let p_{ij} be the probability that an observation lies in the $(ij)^{th}$ cell. If no cell frequencies are missing, then under the null hypothesis of no association between rows and columns

$$p_{ij} = p_i \, q_j \ , \ i = 1,\ldots,r_j \ j=1,\ldots, \ c, \qquad (3.1)$$

where the p_i's (q_j's) are the probabilities of an observation falling into the i^{th} row (j^{th} column), $\sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1.$

If n_{11} is missing, then under the null hypothesis the joint distribution of the available n_{ij}'s is given by an $(rc-1)$ - nomial distribution with cell probabilities

$$p'_{ij} = \frac{p_i q_j}{1-p_1 q_1} \ , \ i =1,\ldots,r; \ j=1,\ldots, \ c; \ (i,j) \neq (1,1) \qquad (3.2)$$

Let N be the total of available frequencies.

The likelihood function for the available frequencies is

$$L = \prod_{(i,j)\neq(1,1)}^{(r,c)} \left(\frac{p_i q_j}{1-p_1 q_1}\right)^{n_{ij}} . \qquad (3.3)$$

Let the prior distribution for the missing frequency n_{11}, $p(n_{11})$ say, be given by the negative binomial distribution

$$p(n_{11}) = \binom{N+n_{11}-1}{n_{11}} (p_1 \, q_1)^{n_{11}} (1-p_1 q_1)^{N}, \; n_{11} = 0,1,2,\ldots \qquad (3.4)$$

The choice of (3.4) for the prior distribution of n_{11} can be motivated by noting that if n_{11} were not missing, then the marginal probability of observing n_{11} in the first cell would equal $\binom{N+n_{11}}{n_{11}}(p_1 q_1)^{n_{11}}(1-p_1 q_1)^{N}$.

We replace $\binom{N+n_{11}}{n_{11}}$ by $\binom{N+n_{11}-1}{n_{11}}$ so that $\sum_{n_{11}=0}^{\infty} p(n_{11}) = 1$.

If the prior distribution of $(p_1,\ldots,p_r,q_1,\ldots,q_c)$ is proportional to

$\prod_{(i,j)=(1,1)}^{(r,c)} (p_i q_j)^{m_{ij}}$, then the posterior distribution of the p_i's, q_j's

and n_{11} is proportional to

$$\binom{N+n_{11}-1}{n_{11}} (p_1q_1)^{n_{11}} (1-p_1q_1)^N \prod_{(i,j)\neq(1,1)}^{(r,c)} \left(\frac{p_i q_j}{1-p_1q_1}\right)^{n_{ij}} \prod_{(i,j)=(1,1)}^{(r,c)} (p_iq_j)^{m_{ij}} . \tag{3.5}$$

In the derivation below we take all m_{ij} equal to zero. Non-zero values can usually be introduced at the end. From (3.5) we than have that the joint posterior distribution of the p_i's and q_j's is given by

$$\text{Post } (p_i\text{'s}, q_j\text{'s}) \propto \sum_{n_{11}=0}^{\infty} \binom{N+n_{11}-1}{n_{11}} (p_1q_1)^{n_{11}} (1-p_1q_1)^N \prod_{(i,j)\neq(1,1)}^{(r,c)} \left(\frac{p_iq_j}{1-p_1q_1}\right)^{n_{ij}} \tag{3.6}$$

$$\propto \prod_{(i,j)\neq(1,1)}^{(r,c)} \left(\frac{p_iq_j}{1-p_1q_1}\right)^{n_{ij}}$$

Jeffreys (see [4]) showed that the joint posterior density (3.6) can, as $N \to \infty$, be approximated by the normal density with exponent $-\frac{1}{2}\ \chi^2$, where

$$\chi^2 = \sum_{(i,j)\neq(1,1)}^{(r,c)} \frac{\left(n_{ij}-Np'_{ij}\right)^2}{Np'_{ij}} . \tag{3.7}$$

χ^2 is an approximation to the likelihood ratio statistic

$$C = -2 \sum_{(i,j)\neq(1,1)}^{(r,c)} n_{ij} \log\left(\frac{N p'_{ij}}{n_{ij}}\right) \tag{3.8}$$

The exact posterior distribution of C is obtainable by using the methods of Watson [7].

The quantity χ^2 has the $\chi^2_{(r\ c)-1}$ distribution as $N \to \infty$. Hence, using theorem 7.5.1 in [5],

$$\tilde{\chi}^2 = \sum_{(i,j)\neq(1,1)}^{(r,c)} \frac{\left(n_{ij} - N \hat{p}'_{ij}\right)^2}{N \hat{p}'_{ij}} \tag{3.9}$$

has the $\chi^2_{(rc-2)-(r+c-2)=(r-1)(c-1)-1}$ distribution as $N \to \infty$. The $\hat{p}'_{ij}$'s are the maximum likelihood estimates for the p_{ij}'s assuming that the null hypothesis of no association between rows and columns is true. Watson [6] showed that the M.L. estimates of the p_i's, q_j's and n_{11} are given by

$$\begin{cases} \hat{p}_1 = \dfrac{R_1+\hat{n}_{11}}{N+\hat{n}_{11}}, \quad \hat{p}_i = \dfrac{R_i}{N+\hat{n}_{11}}, \quad (i = 2, \ldots, r) \\[2ex] \hat{q}_1 = \dfrac{C_1+\hat{n}_{11}}{N+\hat{n}_{11}}, \quad \hat{q}_j = \dfrac{C_j}{N+\hat{n}_{11}}, \quad (j = 2, \ldots, c) \\[2ex] \hat{n}_{11} = \dfrac{R_1 C_1}{N-R_1-C_1} = \dfrac{N \hat{p}_1 \hat{q}_1}{1-\hat{p}_1\hat{q}_1} \end{cases} \tag{3.10}$$

where R_i $(i=1,\ldots,r)$ and $C_j(j=1,\ldots,c)$ are the existing row and column totals. From (3.10) we find that the $\hat{p}'_{ij}$'s are given by

$$\hat{p}'_{1j} = \frac{R_1 C_j}{N(N-C_1)} \;, \quad (j=2,\ldots,c)$$

$$\hat{p}'_{i1} = \frac{C_1 R_i}{N(N-R_1)} \;, \quad (i=2,\ldots,r) \qquad (3.11)$$

$$\hat{p}'_{ij} = \frac{R_i C_j (N-R_1-C_1)}{N(N-R_1)(N-C_1)} \;, \quad (i=2,\ldots,r\,;j=2,\ldots,c)$$

The test given by (3.9) is, operationally, the same as the sampling result.

The joint distribution of $(p_1,\ldots,p_r;q_1,\ldots,q_c;n_{11})$ can be rewritten as

$$\text{Post }(p_i\text{'s},q_j\text{'s},n_{11}) \propto \binom{N+n_{11}-1}{n_{11}} p_1^{R_1+n_{11}} q_1^{C_1+n_{11}} \prod_{i=2}^{r} p_i^{R_i} \prod_{j=2}^{c} q_j^{C_j} \qquad (3.12)$$

From the normalizing constant of the Dirichlet distribution we know that

$$\int_{S_1} p_1^{R_1+n_{11}} \prod_{i=2}^{r} p_i^{R_i}\, d\, p_1 \ldots d\, p_r = \prod_{i=2}^{r} \frac{R_i!(n_{11}+R_1)!}{(N+n_{11}+r-1)!} \quad ;$$

where $S_1 = \left\{ \text{all } p_i \text{ in } (0,1), \sum_{i=1}^{r} p_i = 1 \right\}$

and

$$\int_{S_2} q_1^{C_1+n_{11}} \prod_{j=2}^{c} q_j^{C_j} dq_1 \ldots dq_c = \prod_{j=2}^{c} \frac{C_j!(n_{11}+C_1)!}{(N+n_{11}+C-1)!} \quad ;$$

where $S_2 = \left\{ \text{all } q_j \text{ in } (0,1), \sum_{i=1}^{c} q_j = 1 \right\}$.

Therefore

$$\text{Post } (n_{11}) \propto \frac{(n_{11}+R_1)!\ (n_{11}+C_1)! \binom{N+n_{11}-1}{n_{11}}}{(N+n_{11}+r-1)!\ (N+n_{11}+c-1)\ !} \quad , \quad n_{11} = 0,1,\ldots \qquad (3.13)$$

We use Barnes'(see [2]) asymptotic series for log $\Gamma(x + h)$,

$$\log \Gamma(x+h) = \log \sqrt{2\pi} + (x+h-\tfrac{1}{2}) \log x - x - \sum_{p=1}^{m} \frac{(-1)^p B_{p+1}(h)}{p\,(p+1)\, x^p} + R_{m+1}(x), \qquad (3.14)$$

where the B_p (h)'s are the Bernouille polynomial and $R_{m+1}(x) = O(x^{-(m+1)})$, to obtain the large sample distribution for the missing frequency n_{11}. We find that log Post (n_{11}) is proportional to

$$(n_{11}+R_1+ \tfrac{1}{2})\log(\hat{n}_{11}+R_1)+(n_{11}+C_1+ \tfrac{1}{2})\log(\hat{n}_{11}+C_1)-(N+n_{11}+r+c- \tfrac{1}{2})\log(N+\hat{n}_{11})$$

$$-\log \Gamma(n_{11}+1)+(N-C_1-R_1-\hat{n}_{11})+\sum_{p=1}^{m}\left\{\frac{(-1)^p}{p(p+1)(N+\hat{n}_{11})^p}\right. \tag{3.15}$$

$$\left.[B_{p+1}(n_{11}-\hat{n}_{11}+r)+B_{p+1}(n_{11}-\hat{n}_{11}+c)-B_{p+1}(n_{11}-\hat{n}_{11}+c)-B_{p+1}(n_{11}-\hat{n}_{11})]\right\}$$

$$-\sum_{p=1}^{m} \frac{(-1)^p}{p(p+1)} B_{p+1}(n_{11}-\hat{n}_{11}+1)\left[\frac{1}{(\hat{n}_{11}+R_1)^p}+\frac{1}{(\hat{n}_{11}+C_1)^p}\right]$$

$$+ O\left\{\max\left((\hat{n}_{11}+R_1)^{-(m+1)}, (\hat{n}_{11}+C_1)^{-(m+1)}\right)\right\}.$$

Ucing the identity (see e.g. [1])

$$B_n(x + h) = \sum_{k=0}^{n} \binom{n}{k} B_k(x)\, h^{n-k} \tag{3.16}$$

$$\log \text{Post}(n_{11}) \propto (n_{11}+R_1+\tfrac{1}{2}) \log(\hat{n}_{11}+R_1)+(n_{11}+C_1+\tfrac{1}{2})\log(\hat{n}_{11}+C_1)$$

$$-(N+n_{11}+r+c-\tfrac{1}{2})\log(N+\hat{n}_{11}) - \log\Gamma(n_{11}+1) + (N-C_1-R_1-\hat{n}_{11})$$

$$+\sum_{p=1}^{m}\sum_{k=0}^{p+1} \frac{(-1)^p \binom{p+1}{k}}{p\,(p+1)} B_k(n_{11}-\hat{n}_{11})\times$$

$$\times \left\{a_k\left(\frac{r^{p+1-k}+c^{p+1-k}}{(N+\hat{n}_{11})^p}\right) \frac{-1}{(\hat{n}_{11}+R_1)^p} - \frac{-1}{(\hat{n}_{11}+C_1)^p}\right\} \qquad (3.17)$$

$$+ 0\left\{\max\left((\hat{n}_{11}+R_1)^{-(m+1)}, (\hat{n}_{11}+C_1)^{-(m+1)}\right)\right\},$$

where

$$a_k = \begin{cases} 1, & k = 0, 1, \ldots, p \\ \frac{1}{2}, & k = p + 1. \end{cases}$$

Hence

$$\text{Post}(n_{11}) \propto \frac{(\hat{n}_{11}+R_1)^{n_{11}} (\hat{n}_{11}+C_1)^{n_{11}} (N+\hat{n}_{11})^{-n_{11}}}{n_{11}\,!}$$ + remainder which goes to zero as the observed frequencies → ∞ .

(3.18)

Using the relationships given by (3.10) we thus have that the large sample posterior distribution of n_{11} is the Poisson distribution with mean $\hat{n}_{11} = \dfrac{N\,\hat{p}_1\hat{q}_1}{1-\hat{p}_1\hat{q}_1}$, i.e.,

$$\text{Post}\,(n_{11}) = \frac{\exp\left\{-\dfrac{N\,\hat{p}_1\hat{q}_1}{1-\hat{p}_1\hat{q}_1}\right\}\left(\dfrac{N\,\hat{p}_1\hat{q}_1}{1-\hat{p}_1\hat{q}_1}\right)^{n_{11}}}{n_{11}\,!}\;, \quad n_{11} = 0,\,1,2,\ldots\;.$$

4. Acknowledgement.

The author is indebted to Professor G. S. Watson for helpful discussions.

References

[1] Abramovitz, M. and Stegun , I. A., Handbook of Mathematical Functions N.B.S., Appl. Math. Ser. 55, U.S. Government Printing Office, Washington, D.C. (1964) pp. 804-810.

[2] Barnes, E.W., "The Theory of the Gamma Function," Messeng. Math. 29 (1899), pp. 64-129.

[3] Cochran, W.G. and G.M. Cox, Experimental Designs, § 3.7, pp. 72-74. New York, John Wiley and Sons, Inc., (1950).

[4] Jeffreys, H., Theory of Probability , Oxford University Press, England,(1939) pp. 145-149.

[5] Lindley, D.V., Probability and Statistics from a Bayesian Point of V View, Vol. 2, Cambridge University Press, England (1965)

[6] Watson, G.S. "Missing and 'Mixed-Up' frequencies in Contingency tables," Biometrics, 12, (1956), pp. 47-50.

[7] Watson, G.S. "Some Bayesian Methods related to χ^2, Johns Hopkins Tech. Rep. No. 34, (1965), The Johns Hopkins University.

[8] Yates, F., "The analysis of replicated experiments when the field results are incomplete," Emp. Jour. Exp. Agr. 1 (1933) pp. 129-142.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Department of Statistics
The Johns Hopkins University
Baltimore, Maryland 21218

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

A BAYESIAN APPROACH TO SOME MISSING VALUE PROBLEMS IN ANOVA AND CONTINGENCY TABLES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(Last name, first name, initial)*

Bloch, Daniel A.

6. REPORT DATE
November, 1966

7a. TOTAL NO. OF PAGES
18

7b. NO. OF REFS
8

8a. CONTRACT OR GRANT NO.
Nonr 4010(09)

b. PROJECT NO.
NR 042-232

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report #62

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. AVAILABILITY/LIMITATION NOTICES

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Logistics and Mathematical Statistics
Office of Naval Research Branch
Washington, D.C.

13. ABSTRACT

Non-Bayesian procedures for dealing with missing values in ANOVA and contingency tables are well-known In this paper we derive the appropriate Bayesian procedures for a randomized block design and for and r x c contingency table. The posterior distributions for the missing observation are given. This paper shows that the classical non-Bayesian procedures do have a simple and natural Bayesian interpretation.

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Bayesian						
missing value						
chi-square						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract o. grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.